Traitement informatique des langues et recherche documentaire

Bernard Victorri

Revue d'Interaction Homme-Machine, 2, 1999 – Version préliminaire

Traitement automatique des langues et recherche documentaire

Bernard Victorri

Introduction

La recherche d'information constitue aujourd'hui l'un des grands enjeux de la communication homme machine. Les grands bouleversements technologiques de ces dernières années, et en premier lieu le développement fulgurant d'Internet, posent un problème redoutable pour qui se préoccupe de la convivialité dans l'utilisation de ces nouvelles ressources. Les techniques classiques de l'informatique documentaire, qui avaient permis de grands progrès dans ce domaine, ne sont plus de mise aujourd'hui. En effet, l'ensemble des documents accessibles à un utilisateur à partir de son ordinateur, ce que l'on peut appeler "l'univers documentaire en ligne", s'est profondément transformé ces dernières années. Le changement est d'abord quantitatif : la masse de documents accessibles s'est accrue et continue de s'accroître à un rythme vertigineux. Mais le changement est aussi qualitatif : alors qu'il y a dix ans, ces documents provenaient essentiellement d'un nombre limité de fournisseurs spécialisés dans l'offre d'informations, aujourd'hui tout le monde met à la disposition de tout le monde des documents de toute sorte, construisant ainsi un univers documentaire à la fois unifié dans son mode d'accès, et complètement éclaté dans la diversité de ses sources.

Du coup, on a pu assister à ce qui ressemble à une régression qualitative des techniques de l'informatique documentaire. Toutes les méthodes qui reposaient sur un travail en amont de la part des producteurs (constitutions de thésaurus, systèmes de descripteurs, travail d'indexation, etc.) ne sont plus adaptées à cette prolifération de documents, et seule la force brute de la recherche "plein texte" combinée à des techniques statistiques de fréquence lexicale sont à l'œuvre dans les moteurs de recherche sur le Web. Tout le monde convient que ces techniques sont insuffisantes, qu'elles produisent trop de bruit (et aussi d'ailleurs de silence!) pour être pleinement efficaces. Mais ce sont pour l'instant les seules qui sont capables de s'attaquer à ce nouvel univers documentaire, et elles constituent une base incontournable : le défi qui se pose à l'informatique documentaire consiste à découvrir comment progresser, à partir de ces techniques, pour améliorer leur taux de réussite dans la sélection de documents pertinents. L'enjeu est décisif : du succès de ces recherches dépend la viabilité de ce nouvel univers documentaire, dont la richesse grandissante risque de devenir de moins en moins exploitable.

Dans le même temps, on a assisté à une autre révolution dans le domaine du traitement automatique des langues. La disponibilité d'énormes corpus et la confection d'imposants dictionnaires électroniques ont, là aussi, changé la donne. On a vu apparaître des techniques radicalement différentes de celles que nous avaient léguées les premières recherches en informatique linguistique. En particulier, la mise en œuvre de méthodes statistiques capables d'exploiter de gros corpus, alliée à l'utilisation judicieuse de connaissances linguistiques appropriées, ont permis de réaliser des analyseurs "robustes", pouvant traiter du texte tout venant, avec des objectifs assez modestes du point de vue linguistique, mais qui correspondent à des besoins bien adaptés à certaines tâches, en particulier pour l'informatique documentaire.

L'objectif de cet article est de montrer que ces outils de traitement automatique des langues peuvent contribuer au renouveau de l'informatique documentaire, en apportant le type de connaissances nécessaires à la résolution des problèmes auquel elle est aujourd'hui confrontée.

1. Les nouveaux enjeux de la recherche documentaire

1.1. Un dispositif centré sur l'utilisateur

Du point de vue de l'utilisateur, la problématique de la recherche documentaire reste fondamentalement la même : il faut sélectionner parmi une grande masse de documents un sous-ensemble de documents "pertinents", c'est-à-dire dont le thème soit en adéquation avec les besoins de l'utilisateur, tels qu'il a pu les exprimer dans sa requête. Ce qui a changé, c'est le rapport aux fournisseurs des informations. D'abord, l'ensemble des fournisseurs n'est plus stable, ce qui fait que l'on ne peut plus déterminer où sont susceptibles de se trouver les documents recherchés. De plus, on ne peut plus compter sur un travail de structuration des ensembles documentaires par les fournisseurs eux-mêmes : cette structuration est souvent purement et simplement inexistante, et, quand elle existe, elle n'est pas la plupart du temps vraiment exploitable. En effet, quand un fournisseur organise les documents qu'il rend disponibles, il le fait en fonction de sa propre "vision" de ce fond documentaire, et cela a peu de chance de correspondre à celle de l'utilisateur.

Prenons un exemple concret : un texte parlant des problèmes posés pour la santé humaine par l'utilisation de farines animales dans l'alimentation des bovins. Ce texte peut intéresser aussi bien des biochimistes, des médecins, des spécialistes de physiologie animale, des industriels de l'agroalimentaire, des chefs d'entreprises de restauration, des responsables politiques, de simples ménagères, etc. Il est clair que ces différentes catégories ne vont pas rechercher ce texte à partir de requêtes identiques. Le vocabulaire utilisé dans les demandes sera radicalement différent. De plus, les autres documents à mettre en relation avec ce texte ne seront pas les mêmes suivant les utilisateurs. Comment pourrait-on demander au fournisseur de ce texte de l'indexer d'une dizaine de manières différentes en pensant à tous ceux qui risquent d'en avoir besoin un jour ? Comment pourrait-il organiser son ensemble documentaire (par des liens hypertexte notamment) en tenant compte de cette variété d'usages potentiels ?

L'informatique documentaire doit donc renoncer à intervenir en amont, du côté du fournisseur, et elle doit plutôt se centrer sur le seul pôle qui garde un minimum de stabilité, à savoir l'utilisateur. Il peut paraître contre intuitif de penser que l'ensemble des utilisateurs, qui grandit encore plus vite que celui des fournisseurs, pourrait constituer un pôle de stabilité. Mais c'est en fait en grande partie le cas. De nombreux utilisateurs mènent des recherches documentaires dans le cadre d'une activité précise et continue (recherche scientifique, veille technologique, etc.). Leurs domaines d'intérêt, la logique de structuration de ces domaines, le vocabulaire utilisé, etc., restent relativement constants. De plus, comme cela correspond à un réel besoin de leur part, ces utilisateurs sont plus à même d'investir une partie de leur temps (ou de leur argent) à confectionner des outils de recherche qui leur soient adaptés, c'est-à-dire qui leur permettent de "voir" l'univers documentaire à partir de leurs propres lunettes.

Ainsi, ce serait en aval, à partir de besoins bien identifiés de groupes d'utilisateurs, que pourrait agir l'informatique documentaire, en offrant à chacun une "grille de lecture" appropriée de la masse foisonnante et exubérante de documents qui peuplent l'univers documentaire[1].

Comme on va maintenant le voir, cela réclame de mettre à la disposition des utilisateurs des outils puissants de traitement de ces documents, dans lesquels les techniques de traitement automatique des langues joueraient un rôle essentiel.

---

[1] Déjà pour l'informatique documentaire traditionnelle, cette nécessité de se centrer sur les utilisateurs était manifeste. C'est dans cette optique, par exemple, qu'ont été conçus des "anté-serveurs", interfaces entre utilisateurs et bases documentaires (cf. [THOM97], [VICT96]).

1.2. La structuration du lexique

Le lexique joue un rôle primordial dans la recherche documentaire. Quand un utilisateur effectue une recherche documentaire, il l'exprime dans des termes qui sont ceux de sa propre spécialité, qui ne correspondent pas forcément à ceux qui sont utilisés dans les documents qu'il recherche. Cela est déjà vrai dans sa propre langue, à cause du phénomène de la synonymie, amplifiée par le fait que ces documents n'ont pas été forcément rédigés par des auteurs de cette spécialité, et l'on sait que les lexiques varient énormément avec les domaines de compétence. C'est encore plus vrai pour des documents rédigés dans une autre langue : l'utilisateur peut très bien comprendre cette langue, mais ne pas être suffisamment familiarisé avec elle pour connaître les termes effectivement utilisés par les spécialistes du domaine. En outre, la polysémie d'un grand nombre d'unités lexicales vient ajouter une difficulté supplémentaire : les termes de l'utilisateur n'ont la signification qu'il leur prête que dans le cadre de son domaine, ce qui explique en grande partie le bruit auxquelles conduisent les recherches "plein texte".

Il faut donc donner les moyens à l'utilisateur de constituer une véritable base de données lexicale, de manière à ce que le système d'interrogation puisse retrouver, à partir de la requête de l'utilisateur, l'ensemble de termes le plus pertinent possible pour effectuer la recherche de documents. Cette base de données lexicale doit être organisée à partir du vocabulaire de l'utilisateur, en fonction de la cohérence de ses besoins : il doit pouvoir réaliser sa propre "ontologie" en structurant la hiérarchie des termes suivant la vision du domaine qui est la sienne. Et, idéalement, tous les termes susceptibles d'être trouvés dans les documents, que ce soit dans sa langue ou dans une autre, doivent figurer dans la base à la place qui leur correspond dans cette ontologie.

Bien entendu, l'utilisateur ne peut pas construire une telle base de données indépendamment des textes qu'il est susceptible de rechercher. Qui plus est, il est très difficile, voire impossible, à un utilisateur de construire ex nihilo une telle base, même en se limitant à son propre vocabulaire : on sait la difficulté que représente l'explicitation des connaissances d'un domaine, même pour des spécialistes qui le maîtrise parfaitement.

Il faut donc concevoir des systèmes d'aide à la constitution de ces bases lexicales. De manière idéale, ces systèmes doivent pouvoir prendre en entrée un corpus de textes, en faire l'analyse, et proposer à l'utilisateur une esquisse de base de données lexicale que l'utilisateur pourra valider et modifier à sa guise. Ces systèmes doivent être évolutifs : à partir de tout nouveau texte qui lui est soumis, le système doit proposer des ajouts à la base de données qui tiennent compte des nouveaux termes rencontrés. A l'utilisateur bien entendu de valider ou non ces ajouts, et d'enrichir ainsi progressivement la base de données lexicale qui constitue les lunettes à travers lesquelles il voit l'univers documentaire.

Il faut noter que de telles bases lexicales ne sont pas destinées à un utilisateur unique. Elles peuvent être partagées par tout un groupe d'utilisateurs qui possèdent les mêmes besoins documentaires, et donc devenir le bien commun d'une communauté, soit qu'elles aient été le fruit d'un travail collectif ou qu'elles aient été réalisées par une entreprise à leur intention ou à leur service.

Contrairement aux apparences, la réalisation de tels systèmes d'aide n'est plus une utopie. Comme on le verra au §2, il existe aujourd'hui des outils de traitement automatique de textes qui permettent d'envisager très concrètement la mise en place de ces systèmes, qui, rappelons-le, ne sont que des systèmes d'aide : il serait illusoire de penser que l'on pourra complètement automatiser la construction de ces bases lexicales. En revanche, l'aide automatique que l'on peut fournir devrait pouvoir rendre cette tâche réalisable en fournissant à l'utilisateur toutes les données qui lui sont nécessaires.

1.3. L'indexation automatique

Le deuxième aspect de la recherche documentaire qui pourrait bénéficier des nouvelles techniques de traitement automatique des langues concerne l'indexation automatique des documents. A l'heure actuelle, les documents sont essentiellement vus comme des suites de mots, et les principales informations qui sont tirées de ces suites sont de nature statistique : essentiellement la fréquence d'un mot, et parfois des relations de proximité entre deux mots.

En gros, de manière à peine caricaturale, un mot est considéré comme d'autant plus représentatif du contenu d'un document qu'il est employé plus fréquemment dans le texte. Or il est clair que la fréquence est loin d'être un critère fiable.

D'abord, un certain nombre de mots très fréquents ne disent rien sur le contenu d'un texte : il s'agit des mots dits "grammaticaux" (articles, prépositions, etc.), mais aussi de certains mots "lexicaux" (noms, verbes, adjectifs) dont le contenu sémantique est trop vague pour être discriminants. Cela est déjà pris en compte dans les systèmes actuels, au moins en partie, soit en écartant ces mots de la requête, soit en diminuant leur importance dans la recherche grâce à un calcul plus fin de la fréquence (en pondérant par exemple la fréquence d'un mot dans le texte par sa fréquence dans la langue en général, calculée à partir d'un vaste ensemble de documents). L'ennui de ces méthodes, c'est que l'unité pertinente pour la recherche n'est pas forcément le mot : c'est le plus souvent un groupe nominal complexe (comportant en particulier des prépositions) dont il faudrait conserver l'intégrité. Par exemple, un texte parlant de *chemin de fer* est très mal indexé par les deux descripteurs indépendants *chemin* et *fer*.

Un autre problème concerne la morphologie grammaticale. Dans les langues à morphologie riche (comportant des marques de genre, de nombre, de cas, etc.), un même mot peut apparaître sous de nombreuses formes différentes, ce qui fausse le calcul de fréquence. Là encore, les systèmes actuels cherchent à prendre en compte ce phénomène, mais avec des méthodes beaucoup trop rudimentaires.

Une autre difficulté provient de l'existence dans toutes les langues du phénomène d'anaphore, qui permet d'éviter de répéter les mots qui constituent le thème du discours. Pour calculer la fréquence réelle d'un terme dans un texte, il faudrait comptabiliser aussi toutes les fois où il a été remplacé par un pronom, ce qui changerait radicalement la représentativité des différents termes, puisque ce sont justement les termes qui constituent le sujet du texte qui sont le plus soumis à ce mécanisme d'anaphore.

Toutes ces difficultés montrent bien que l'on ne peut pas faire l'impasse sur certains éléments d'analyse linguistique des textes si l'on veut améliorer la recherche documentaire. Comme on va le voir, on dispose là aussi d'outils de traitements automatiques des langues capables de mener à bien ces tâches d'analyse, pas de manière totalement fiable, certes, mais avec un taux de réussite suffisant pour améliorer substantiellement l'indexation des documents. Mieux, comme on le verra, cette indexation peut aussi se faire à un niveau plus fin que celui du texte : on peut repérer des parties de document qui présentent une certaine unité thématique, et donc caractériser ces parties par un ensemble de descripteurs, ce qui permettrait de mieux cibler encore la sélection de textes pertinents.

Il faut noter aussi que la mise en place de ces outils de traitement linguistique de textes, qui réclament bien sûr des temps de calcul non négligeables, ne représente plus aujourd'hui une difficulté technique insurmontable. L'utilisation d'"agents" qui visitent les sites de manière systématique pour construire des systèmes d'indexation est déjà largement répandue sur Internet. On peut raisonnablement penser que les développements technologiques permettront d'aller plus loin

dans cette voie, et que des traitements de plus en plus lourds seront rendus possibles sur de grandes masses de documents.

2. Les nouveaux outils de traitement automatique des textes

2.1. L'analyse morpho-syntagmatique

Traditionnellement, les deux premières étapes de l'analyse linguistique sont l'analyse morphologique et l'analyse syntaxique (cf. [FUCH93]). L'analyse morphologique consiste à déterminer la catégorie syntaxique (nom, verbe, etc.) de chaque mot et sa lemmatisation (*finir* pour *finissaient*). L'analyse syntaxique consiste à construire un arbre qui représente l'organisation hiérarchique de la phrase en syntagmes récursifs.

Le traitement automatique a longtemps buté sur une difficulté qui était due à ce découpage en deux étapes. En effet, d'une part, il est impossible de mener à bien l'analyse morphologique sans faire appel à des connaissances syntaxiques (dans la phrase *la préposée mesure la pression*, il faut analyser toute la phrase pour attribuer à *mesure* la catégorie verbale), et d'autre part, on ne peut pas obtenir une analyse syntaxique complète sans connaissances sémantiques (comparer *réparer le câble au tungstène* et *réparer le câble au chalumeau*).

Cette difficulté a été levée ces dernières années avec l'apparition d'une nouvelle génération d'analyseurs qui réalisent en même temps l'analyse morphologique et une première étape de l'analyse syntaxique consistant à segmenter la phrase en *chunks* [ABNE92] ou syntagmes non-récursifs. Sans entrer dans les détails ici (on trouvera une description de cette approche dans [VERG98]), indiquons simplement que cette segmentation ne cherche pas à résoudre les problèmes de rattachement de groupes prépositionnels (ainsi *réparer le câble au chalumeau* est segmenté en trois groupes *réparer*, *le câble* et *au chalumeau* sans rattacher le troisième groupe à l'un des deux premiers. En revanche, cette méthode permet de réaliser avec beaucoup de sûreté l'analyse morphologique, et elle est très robuste, puisqu'elle peut traiter avec un excellent taux de succès des textes tout-venant, même avec un dictionnaire incomplet.

Ce type d'analyse est particulièrement bien adapté à l'informatique documentaire. En effet, ce qui importe en premier lieu pour la recherche documentaire, c'est la reconnaissance des syntagmes nominaux : on n'a pas besoin d'une analyse syntaxique complète. Bien entendu, les termes pertinents sont en général des syntagmes nominaux complexes, comme *câble au tungstène*. Comme on va le voir ci-dessous, d'autres techniques doivent être utilisées pour déterminer ces syntagmes complexes. Mais cela ne peut se faire que si l'on dispose déjà de cette première segmentation en syntagmes non-récursifs, qui permet de détecter tous les groupes nominaux, et bien sûr, d'obtenir les formes lemmatisées des noms et adjectifs qui les composent.

Il faut noter que cette étape d'analyse ne réclame que des moyens informatiques relativement légers : ce sont des traitements très rapides, et, comme on l'a dit, ils ne nécessitent pas un dictionnaire exhaustif. Ils sont donc utilisables, comme première étape de traitement, aussi bien pour la tâche de constitution de bases de données lexicales (§1.2), que pour la tâche d'indexation (§1.3).

2.2. L'extraction de syntagmes nominaux pertinents

Une fois cette première analyse réalisée, le problème essentiel qui se pose à l'informatique documentaire consiste donc à extraire les syntagmes nominaux complexes qui sont susceptibles de

représenter une notion pertinente pour le domaine, et qui devront donc être intégrés dans la base de données lexicale et servir à l'expression des requêtes. Pour reprendre notre exemple, il s'agit de différencier *câble au tungstène*, qui est un candidat plausible, de *câble au chalumeau*, qui n'est pas recevable, puisque *au chalumeau* n'est pas, en fait, rattaché à *câble*.

Pendant longtemps, on a cru que ce problème ne pouvait être résolu qu'en faisant intervenir des connaissances sémantiques, et qu'il était donc impensable de traiter correctement des textes tout-venant de ce point de vue, puisque cela impliquerait de réaliser d'abord d'énormes bases de connaissances des domaines traités, bien plus puissantes que les simples bases lexicales que l'on cherche à construire. Mais là encore, la situation est en train de changer radicalement.

En effet, depuis que l'on peut travailler efficacement sur de très gros corpus, de nouvelles techniques, que l'on peut appeler "statistico-linguistiques", permettent d'aborder ce problème par un autre biais. La méthode consiste à réaliser sur ces gros corpus des analyses distributionnelles[2], de manière à relever tous les contextes dans lesquels apparaît un groupe nominal donné, et d'en déduire si un rattachement est correct ou non. Nous n'entrerons pas dans les détails techniques assez complexes de cette méthode (cf. [HABE96, NAUL98]), mais dans le principe, l'idée est assez simple, comme on va le voir sur notre exemple. Si l'on dispose d'un corpus suffisamment important, on va trouver des occurrences de *au chalumeau* dans lesquelles ce groupe nominal prépositionnel suit directement un verbe de la même "classe" que *réparer* (comme dans *Cette plaque doit être découpée au chalumeau*). On peut alors en déduire que dans *réparer un câble au chalumeau*, le groupe *au chalumeau* se rattache plutôt à *réparer* qu'à *câble*. En revanche, si l'on examine toutes les occurrences de *au tungstène*, on le trouvera systématiquement derrière des noms de la même "classe" que *câble*, ce qui permettra de faire là aussi le bon rattachement. Reste à définir ce que l'on entend par "classe". Il s'agit bien sûr, idéalement, de classes sémantiques. Mais tout l'intérêt de la méthode est de ne pas définir ces classes a priori : elles vont émerger elles-mêmes de l'analyse distributionnelles. En effet, on associe à chaque terme l'ensemble de ses contextes, et l'on regroupe les termes qui ont suffisamment de contextes identiques dans une même classe. Le processus est itératif : au fur et à mesure que des termes sont regroupés dans ces classes, les contextes qu'ils forment pour d'autres termes sont considérés comme équivalents, ce qui permet de regrouper à leur tour certains de ces termes.

Ainsi, ces techniques permettent de réaliser deux tâches : d'une part, extraire les syntagmes nominaux complexes qui sont des candidats descripteurs du document, et d'autre part, regrouper ces termes en classes susceptibles de constituer un premier niveau de structuration de la base de données lexicale que l'on cherche à construire. On possède donc là des outils tout à fait adaptés aux besoins de l'informatique documentaire que nous avons présentés ci-dessus.

Pour que ces méthodes fonctionnent correctement, il faut impérativement disposer de gros corpus. Mais de toute façon, quelle que soit la taille du corpus, elle ne sont pas fiables à 100%. C'est pourquoi elles ne peuvent être utilisées que dans le cadre d'une aide à la constitution de bases de données lexicales : il faut l'intervention de l'utilisateur pour valider ou invalider les regroupements, et construire de manière cohérente la hiérarchie des termes du domaine. Mais bien sûr l'utilisateur est puissamment aidé dans cette tâche, puisqu'il dispose d'un stock de termes candidats, déjà grossièrement classés, pour accomplir sa tâche. De tels systèmes d'aide commencent à voir le jour : à titre d'exemple, on peut citer le travail d'Elie Naulleau, qui a réalisé un tel système et qui l'a testé sur un gros corpus de documentation de l'EDF.

[2] L'analyse distributionnelle a une longue histoire en linguistique (cf. [HARR71]. Mais ce n'est que très récemment que les progrès technologiques ont rendu son application possible au traitement automatique des langues.

2.3. Les traitements thématiques

On a assisté aussi ces dernières années à d'autres progrès du traitement automatique des langues qui sont du plus grand intérêt pour l'informatique documentaire : il s'agit de ce que l'on peut appeler l'analyse "discursive" des textes. En effet, un texte n'est pas constitué d'une suite de phrases isolées, indépendantes les unes des autres. Il existe des relations entre ces phrases, qui donnent à l'ensemble sa cohérence. Cette cohérence est due au fait qu'un texte aborde généralement tour à tour un certain nombre de points, que l'on appelle des "thèmes". Certains thèmes peuvent couvrir l'ensemble du texte, et d'autres se limiter à une partie plus ou moins importante. La détermination de ces thèmes est capitale pour la recherche documentaire puisque le but de l'indexation est justement de représenter par quelques termes ce dont parle le texte.

Il existe déjà des techniques pour déterminer les thèmes d'un document, en s'appuyant sur la structure du document : titres, sous-titres, résumé, etc. Ces techniques sont très efficaces et très sûres, et elles méritent d'être plus développées et davantage utilisées en informatique documentaire. Nous ne nous attarderons pas là-dessus ici, mais nous allons plutôt présenter d'autres méthodes, purement linguistiques, qui s'attaquent au texte lui-même, et qui peuvent donc jouer un rôle complémentaire dans cette recherche des thèmes, en particulier dans les cas où les documents sont très peu structurés.

Il existe deux types de marques linguistiques qui peuvent être utilisées pour la recherche des thèmes : les marques anaphoriques et les connecteurs discursifs.

Le phénomène de l'anaphore représente aujourd'hui un handicap pour l'indexation automatique, comme on l'a vu au §1.3. Mais, si l'on sait le traiter, cela peut devenir au contraire un moyen très efficace de déterminer les thèmes. En effet, la résolution des anaphores conduit à construire ce que l'on appelle des "chaînes de coréférence", constituées d'une suite de groupes nominaux et de pronoms qui réfèrent à la même entité dans le discours. Plus on parle d'une entité, plus la chaîne de coréférence qui lui est associée est longue : c'est donc un indice très sûr pour déterminer l'importance d'un thème.

Or on dispose aujourd'hui de bonnes méthodes pour résoudre les anaphores. Les anciennes techniques très rudimentaires, qui consistaient en gros à prendre comme antécédent d'un pronom le nom le plus proche de même genre et de même nombre, ont laissé la place à des méthodes qui utilisent une notion de "saillance" (cf. [ALSH87], [ARIE90], [DUPO97]). La saillance d'une entité est d'autant plus grande que le lecteur peut s'attendre à rencontrer une occurrence de cette entité dans la suite du texte. La saillance dépend de plusieurs facteurs : le nombre de fois que l'on a déjà parlé de cette entité, la proximité de la dernière occurrence (la "récence"), la nature des termes utilisés pour en parler, etc. La plupart de ces facteurs sont facilement calculables, et le système peut donc maintenir en mémoire le degré de saillance des entités déjà rencontrées. Chaque fois que l'on rencontre une anaphore, on utilise ce degré de saillance, ainsi que d'autres critères complémentaires, pour résoudre l'anaphore et remettre à jour les degrés de saillance.

Ainsi, non seulement on peut résoudre les anaphores, mais on peut aussi disposer directement, grâce à la saillance, d'une mesure de l'importance thématique de chaque entité. De plus, on peut se servir de ces traitements pour détecter l'unité thématique des différentes parties du texte : un changement de thème correspond à un arrêt des chaînes de coréférence, et à un changement important de la saillance des entités.

L'analyse des connecteurs discursifs peut aussi contribuer à déterminer la structure thématique du texte. Les connecteurs discursifs sont des expressions de la langue, comme *Par conséquent*, *Par ailleurs*, *D'un autre point de vue*, *Selon M. X*, etc., qui servent justement à indiquer le statut thématique de ce qui suit par rapport à ce qui précède. Un certain nombre de travaux en traitement

automatique des langues, en particulier sur le résumé automatique (cf. [DESC97]), ont montré que la prise en compte de ces connecteurs était très utile pour déterminer les changements thématiques et, d'une manière générale, la structure thématique d'un document.

Conclusion

Ainsi les progrès du traitement automatique des langues permettent d'envisager de manière réaliste leur mise en œuvre à grande échelle en informatique documentaire. Bien sûr, les méthodes présentées ici sont encore, pour la plupart, à l'état de recherche. Beaucoup de travail est encore nécessaire avant que ces techniques deviennent complètement opérationnelles. Mais que se soit pour constituer des bases de données lexicales à partir de corpus, ou pour indexer automatiquement des documents, on devrait disposer dans un avenir d'outils adaptés, capables de travailler sur des documents tout-venant et de mettre au service de l'informatique documentaire un certain nombre des connaissances linguistiques, bien ciblées, qui peuvent améliorer de manière spectaculaire les performances de la recherche documentaire.

Bibliographie

[ABNE92]
Abney S., "Prosodic structure, performance structure and phrase structure", *Proceedings of Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.

[ALSH87]
Alshawi H., *Memory and context for language interpretation*, Cambridge University Press, 1987

[ARIE90]
Ariel M., *Accessing Noun Phrases Antecedents*, London, Routledge, 1990.

[DESC97]
Desclés J.P., "Systèmes d'exploration contextuelle", in Guimier C. (ed.), *Co-texte et calcul du sens*, Presses Universitaires de Caen, 1997.

[DUPO97]
Dupont M., Clavier S., "Calcul des Chaînes de Référence dans des Textes Tout Venant", *Actes de JST'96*, 1997, p.345-351.

[FUCH93]
Fuchs C. *et al.*, *Linguistique et Traitements Automatiques des Langues*, Hachette, 1993.

[HABE96]
Habert B., Fabre C., "Simplifying nominal parse trees to find semantic types in corpus", in Ide N. (ed.), *Research in Humanities*, Kluwer, 1996.

[HARR71]
Harris Z., *Structures mathématiques du langage*, Dunod, 1971.

[NAUL98]
Naulleau E., *Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire*, Thèse de l'Université de Paris 13, 1998.

[THOM97]
Thomazo L., *L'anté-serveur documentaire*, Thèse de l'Université de Caen, 1997.

[VERG98]
Vergne J., Giguet E., "Regards théoriques sur le *tagging*", *Actes de la cinquième conférence sur le Traitement automatique des Langues Naturelles (TALN'98)*, 1998.

[VICT96]
Victorri B., "Les systèmes informatiques doivent-ils modéliser leurs utilisateurs ?", in Vivier J. (ed.), *Psychologie du dialogue homme-machine en langage naturel*, Europia Productions, 1996.